DOCUMENT RESUME

ED 312 297 TM 014 069

AUTHOR Lawson, Stephen
TITLE Latent Trait Measurement: Do the Results Justify the Effort?
PUB DATE Nov 89
NOTE 18p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Correlation; Estimation (Mathematics); *Latent Trait Theory; Mathematical Models; *Measurement; Theory Practice Relationship; Validity
IDENTIFIERS Data Sets; *Rasch Model

ABSTRACT

During the past several decades, a new class of measurement theory, called latent trait (or item response) theory, has become popular. Supporters contend that results obtained through the use of this theory prove vastly superior to classical measurement statistics. This paper investigates the use of the Rasch model, one of the more popular latent trait models, and compares results with those obtained through classical measurement methods. Three data sets consisting of dichotomously scored data from different tests were analyzed. Data Set 1 consisted of responses from 35 subjects across 18 test items. Data Sets 2 and 3 consisted of responses to 35 test items, with 791 and 797 subjects in Data Sets 2 and 3, respectively. These sets of data were used to calculate product-moment correlations between estimates derived from both approaches. Results indicate that Rasch statistics are similar to their classical measurement theory counterparts, suggesting that the more sophisticated model may not always be the preferred model. Three tables present data set information. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED312297

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEPHEN LAWSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LATENT TRAIT MEASUREMENT:

DO THE RESULTS JUSTIFY THE EFFORT?

Stephen Lawson

University of New Orleans

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 8-10, 1989.

TM014069

BEST COPY AVAILABLE

ABSTRACT

During the past several decades, a whole new class of measurement theory, called latent trait (or item response) theory, has become popular. Supporters contend that results obtained through the use of this theory prove vastly superior to classical measurement statistics. This paper investigates the use of the Rasch model, one of the more popular latent trait models, and compares results with those obtained through classical measurement methods. Several data sets are used to calculate product-moment correlations between estimates derived from both approaches. It is ascertained that Rasch statistics are little different that their classical measurement theory counterparts, suggesting that the more sophisticated model may not always be the preferred model.

Classical measurement theory has traditionally held that an individual's test score is composed of two independent components: a "true score" and an "error component" (Thorndike, 1982). The true score component refers to the examinee's innate ability on the particular scale of interest. It is this latent ability that we attempt to measure through the use of various test items. The true score, as opposed to the actual test score, remains constant from one occasion to another and from one form of a test to a parallel form. The second component of a test score, the error component, reflects the varying degree to which the obtained test score differs from the true score.

Because an individual's true score cannot be directly observed and measured, it must be inferred from data obtained on responses to various test items. Measurement specialists have traditionally relied on the use of three statistical indices for analyzing and interpreting test results. Perhaps the most familiar test statistic is the individual's total test score, consisting of the number of correct test responses. The total test score provides an estimate of a person's ability. Item P-values reflect the relative difficulty of test items, and are calculated as the proportion of individuals correctly responding to each item. The greater the proportion of correct responses, the easier the item for this sample of individuals. Finally, the power of each item to discriminate between individuals of lesser ability and those of greater ability is also computed. Scores on each item are correlated with the scores on all of the items as a complete set,

or with all the items in the item pool except the given item for which a discrimination coefficient is being computed at one point in time. The higher the correlation, the more useful the item is in providing information regarding a person's ability.

While accepted as standard measurement procedures for the past 40 to 50 years, these procedures have not been without substantial criticism. The most obvious weakness of classical measurement theory centers around the fact that resulting statistics are not invariant or stable across groups of differing ability (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978). The resulting measurements are solely dependent on the group of individuals on whom the statistics were originally obtained.

McKinley and Mills (1989, p. 72), using a 20-item test of mathematics knowledge as an example, explain:

> Typically, an examinee is assigned a score which is a function of the number of items answered correctly by the examinee. Items on the test are assigned proportion-correct item difficulty statistics based on the proportion of examinees responding to each item. If the test is readministered to two groups of different average math skills, or if the test is readministered to the same group after learning takes place, the proportion-correct item difficulties will take on different values. Likewise, if an examinee is administered two tests of different average difficulty, two different scores will result.

If total correct scores and item difficulty values vary from one group to another, one can safely assume that item-total correlations reflecting item discrimination will also vary when obtained from two separate groups taking the same examination.

This measurement instability represents a sizable problem in the design and use of tests. Standard item statistics have therefore been viewed as useful only when the current group of examinees is extremely similar to the sample of examinees from which the item statistics were originally obtained (Hambleton et al., 1978). Wright and Stone (1978), in the forward to their book on best test design, sum up the perceived severity of the situation as follows: "...I think you will find that this small point (of sample dependency of classical measurement statistics) is a matter of life and death to the science of mental measurement. The truth is that the so-called measurements that we now make in educational testing are no damn good!" (p. xi).

In response, test designers have turned to alternative procedures for the design and use of more accurate and reliable tests. One such theory is latent trait theory (also known as item response theory). Latent trait procedures are declared to be "(o)ne of the most important methodological advances in psychological measurement in the past half-century" (McKinley & Mills, 1989, p. 71). Hambleton and Cook (1977) allege that "latent trait models are now widely recognized as desirable alternatives to the classical test model for the design and interpretation of many types of mental tests" and that the use of latent trait

procedures will become "increasingly important to test practitioners" (p. 76).

Advocates of latent trait theories maintain that resulting person and item statistics are vastly superior to their classical measurement counterparts. Latent trait theory states that the probability of an individual's correctly responding to a test item is a function of only two characteristics: the person's ability and the difficulty of the item (Rentz & Rentz, 1978). Latent trait statistics are alleged to estimate an individual's ability independently of his/her test responses. In like manner, item statistics obtained through latent trait calculations are said to be independent of the examinee responses from which item calibrations were derived. Latent trait measurements, therefore, reflect only the true relationship between a person's performance on the test and his or her innate ability. Thus, the primary argument for the use of latent trait models over classical measurement procedures is that latent trait analyses should result in both person free and sample free measurements (McKinley & Mills, 1989). Using latent trait methods, an individual should theoretically receive the same estimate or score of ability regardless of the difficulty of the test given. Statistics regarding item difficulty and item discrimination will also remain stable across different groups of individuals.

This paper investigates the use of the Rasch model, one of the more popular latent trait models, and compares test results with those obtained through classical measurement method. Several data

sets of differing size were analyzed using both methods to determine the degree of difference between the two models, and to evaluate the relative benefits, if any, of Rasch latent trait measurement.

Procedure

Three data sets consisting of dichotomously scored data from different tests were analyzed. With dichotomous scoring, an individual receives a "1" for a correct response or a "0" for an incorrect response. While other latent trait models have been developed to analyze polychotomous scored data (i.e., Likert scale data), usual Rasch model applications assume the data to be dichotomous (McKinley & Mills, 1989).

The data sets have been arbitrarily labelled "Data Set 1", "Data Set 2", and "Data Set 3". Data Set 1 (Wright & Stone, 1979, p. 31) was the smallest of the three sets, consisting of responses from 35 individuals across 18 test items. Data Sets 2 and 3 provided examples of analysis of larger data sets collected using a measure described by Melancon and Thompson (1989). Both of these sets contained 35 test items, with 791 individuals responding to Data Set 2 and 797 individuals responding to Data Set 3.

Standard test and item statistics (i.e., total test scores, item p-values, and item-total correlations) were obtained for all three data sets. Results were computed through the SPSS-X Descriptives and Reliability (Norjusis, 1988). Corresponding Rasch statistics (i.e., person ability estimates, item difficulty estimates, and mean square residuals) were obtained through the use

of BICAL (Wright & Mead, 1976), a computer package designed to provide Rasch model analysis.

In the first step of Rasch analysis, items and persons with all correct or incorrect responses are removed prior to the initial analysis. Because of zero variance of these responses, it is reasoned that "no definite estimates of difficulty or ability can be made" (Wright & Stone, 1979, p. 30). Four items and one respondent from Data Set 1 were removed. Three persons had perfect scores in Data Set 2 and were excluded from the analysis on this basis. No items or persons were removed from Data Set 3.

Latent trait theory holds that a mathematical function can be used to describe the relationship between a test item and a person's ability. Rasch procedures therefore begin by estimating the parameters of this formula or model and estimating fit of data to the model. This process of estimating item and person parameters and evaluating fit to the model is referred to as calibration (Rentz & Rentz, 1978). During calibration procedures, items and persons are again deleted from the analysis, this time when patterns of responses do not fit the estimated model. Final analysis of the three data sets resulted in the following calibrations: Data Set 1 - 14 items calibrated on 32 persons; Data Set 2 - 35 items calibrated on 766 persons; Data Set 3 - 35 items calibrated on 773 persons. All remaining analyses across the three data sets and for both models employed these same item and subject pools.

Results

Results of both classical analysis procedures and Rasch procedures for the three data sets are presented in Tables 1 through 3. The first two columns of each table represent estimates of individual abilities as reflected by item responses. Column 1 presents all possible scores that individuals could have received on the test. The greater the number of correct responses given on the test, the higher the total test score, and thus the higher level of ability that is inferred. Column 2 presents person ability estimates provided through Rasch procedures. The mathematical units used to define both ability estimates and item difficulty estimates are called "logits". Average ability is centered at approximately zero logits.

INSERT TABLES 1 THROUGH 3 ABOUT HERE.

Column 3 indicates the item numbers from the original item pool that were used to provide estimates of both item difficulty and item discrimination as represented in Columns 4 and 5. The higher an item's P-value, the easier that the item is inferred to be. Item difficulty estimates provided through Rasch procedures are centered on an average estimate of zero logits. The easier that the item is estimated to be, the lower the logit value that is assigned to that item.

Columns 6 and 7 represent estimates of each item's ability to discriminate between ability levels of examinees. With classical measurement procedures, this estimate is represented by the item-

total correlation. With Rasch procedures, the corresponding statistic is the item's mean square residual.

Pearson product-moment correlations were calculated between results obtained from each method for all three data sets to investigate degree of similarity among item statistics for both classical and latent trait results. Correlation coefficients are listed at the bottom of each table.

Discussion

The correlation coefficients indicate remarkable similarities between results obtained through classical measurement methods and those results obtained through latent trait methods. Both procedures yield almost identical information regarding both person abilities and item difficulties, usually the two most emphasized estimates in measurement analysis.

Results on item discrimination, however, did differ across data sets. Data Set Three provided the highest correlation ($r = -.9522$) between the two methods. Data Sets One and Two provided lower coefficients ($r = -.7030$ and $r = -.8901$) with respect to item discrimination results. The relatively low correlation coefficient obtained from Data Set 1 may have been due to the small sample size as compared to the other two data sets. Further investigation, using a variety of sizes of data sets, is recommended.

Test and item statistics for several data sets, calculated with both classical measurement procedures and with the more powerful latent trait procedures (specifically, Rasch model procedures), were found to be remarkably similar. Proponents hold

that results obtained through Rasch model procedures are superior to classical measurement estimates because of the unique ability of latent trait models to yield both sample-free and item-free statistics. Yet, person ability estimates and item difficulty estimates were found to be almost identical from one method to the other.

This finding is not meant to discredit the applicability of Rasch model procedures. Latent trait theory does provide several sensitive statistical procedures which may be useful in some specialized research situations (McKinley & Mills, 1989). Person-fit statistics, whereby individuals whose response patterns deviate significantly from the expected pattern, and item-fit statistics, whereby item response patterns can be examined to determine which do not fit the expected pattern, can provide important information in the development of instruments.

In addition, latent trait models allow for the application of true individualized or tailored testing (Hambleton et al., 1978). Individualized testing involves the use of a computer for selecting test items that elicit the greatest amount of information from a specific individual. The person's ability is first estimated by the examiner and an appropriate item is given. If the person correctly responds, his/her ability is recalculated using this information and a more difficult item is given. If the person responds incorrectly, the following item will be of lesser difficulty. Items are given in this manner with the person's ability level re-estimated after each response. Individualized

testing can therefore provide the greatest amount of information regarding a person's ability in the shortest possible time span, since such testing focuses at each test-taker's ability.

However, for the majority of teachers, assessment specialists, and others involved in the design and use of testing instruments, the complicated procedures of latent trait models appear to hold few advantages over standard test methodology. Robert L. Thorndike (1982, p. 12) may have correctly summed up the future of latent trait methods in his opening address at an Australian Council for Educational Research conference focusing on these models:

> For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will a be great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties.

In this case, one must ask, "Are latent trait methods really worth the effort?"

References

Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14(2), 75-96.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48(4), 467-510.

McKinley, R.L., & Mills, C.N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1), (pp. 71-135). Greenwhich, CT: JAI Press Inc.

Melancon, J.G., & Thompson, B. (1989). Measurement characteristics of the Finding Embedded Figures Test. Psychology in the Schools, 26, 69-78.

Norjusis, M. (1988). SPSS-X user's guide (3rd ed). Chicago: SPSS, Inc.

Rentz, R.R., & Rentz, C.C. (1978). Does the Rasch model really work? A discussion for practitioners. Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations. (ERIC Document Reproduction Service No. ED 169 137)

Thorndike, R.L. (1982). Educational measurement: Theory and practice. In D. Spearritt (Ed.), The improvement of measurement in education and psychology: Contributions of latent trait theory (pp. 3-13). Princeton, NJ: ERIC Clearinghouse of Tests,

Measurements, and Evaluations. (ERIC Document Reproduction Service No. ED 222 545)

Wright, B.D., & Mead, R.J. (1976). BICAL: Calibrating items with the Rasch model (Research memorandum No. 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wright, B.D., & Stone, M.H. (1979). Best test design. Chicago: Mesa Press.

Table 1
TEST AND ITEM STATISTICS FOR DATA SET 1 (n = 35-3=32)

CLASS Total Test Score	RASCH Person Ability Estimate	Item No.	CLASS Item P Value	RASCH Item Diff. Estimate	CLASS Item Total Corr.	RASCH Mean Square Residual
1	-6.00	4	93.8%	-4.847	.3131	1.12
2	-5.03	5	90.7%	-4.244	.3205	1.29
3	-4.28	6	90.7%	-4.244	.3759	1.18
4	-3.57	7	96.9%	-5.708	.3420	0.91
5	-2.76	8	81.3%	-2.879	.6755	0.45
6	-1.70	9	87.5%	-3.742	.5447	0.74
7	-0.07	10	68.8%	-1.483	.4038	1.41
8	1.51	11	31.3%	1.511	.4134	1.11
9	2.60	12	15.6%	3.012	.3035	1.18
10	3.53	13	21.9%	2.348	.4702	0.79
11	4.37	14	6.3%	4.435	.1381	1.57
12	5.19	15	3.1%	5.280	.2650	0.98
13	6.16	16	3.1%	5.280	.2650	0.98
		17	3.1%	5.280	.2650	0.98
	r=.9918		**r=.-9949**		**r=.-7030**	

Note: Total Test Scores are based on a possible maximum of 13 points. Four of the original 18 items were removed prior to analysis due to zero variance. Individuals responding correctly to all items or responding incorrectly to all items were also removed during analysis.

Table 2
TEST AND ITEM STATISTICS FOR DATA SET 2 (n = 788-22=766)

CLASS	RASCH	Item	CLASS	RASCH	CLASS	RASCH
Total Test Score	Person Ability Estimate	No.	Item P Value	Item Diff. Estimate	Item Total Corr.	Mean Square Residual
1	-3.94	1	94.5%	-2.121	.1698	0.99
2	-3.21	2	80.9%	-0.636	.2732	0.96
3	-2.75	3	88.8%	-1.311	.2309	0.98
4	-2.41	4	73.7%	-0.186	.2381	1.01
5	-2.12	5	40.1%	1.396	.2397	1.01
6	-1.88	6	60.5%	0.493	.3122	0.97
7	-1.67	7	20.7%	2.495	.1062	1.09
8	-1.47	8	67.7%	0.143	.3208	0.95
9	-1.29	9	33.2%	1.771	.2008	1.05
10	-1.12	10	31.9%	1.838	.0739	1.16
11	-0.96	11	47.6%	1.084	.2439	1.02
12	-0.80	12	73.7%	-0.186	.3448	0.93
13	-0.65	13	55.1%	0.740	.2769	0.99
14	-0.51	14	82.1%	-0.720	.1873	1.02
15	-0.36	15	37.6%	1.552	.3165	0.95
16	-0.22	16	74.4%	-0.224	.2919	0.96
17	-0.08	17	53.7%	0.806	.2510	1.02
18	0.06	18	91.4%	-1.619	.2733	0.93
19	0.20	19	86.0%	-1.034	.2421	0.97
20	0.34	20	46.2%	1.149	.2540	1.01
21	0.49	21	88.6%	-1.283	.2742	0.94
22	0.63	22	87.1%	-1.141	.2286	0.97
23	0.79	23	94.1%	-2.046	.2737	0.92
24	0.94	24	63.7%	0.337	.2235	1.03
25	1.11	25	79.7%	-0.556	.3060	0.95
26	1.28	26	68.7%	0.089	.2543	1.00
27	1.47	27	79.6%	-0.547	.3156	0.94
28	1.67	28	81.8%	-0.701	.3462	0.91
29	1.89	29	68.9%	0.076	.3091	0.96
30	2.13	30	58.8%	0.572	.1340	1.11
31	2.42	31	61.8%	0.431	.2987	0.97
32	2.77	32	69.8%	0.028	.3447	0.93
33	3.24	33	70.6%	-0.014	.3154	0.95
34	3.99	34	66.8%	0.182	.3109	0.96
		35	83.9%	-0.859	.1997	1.01
r=.9860			r=.-9855		r=.-8901	

Note: Total Test Scores are based on a possible maximum of 34 points. Individuals responding correctly to all items or responding incorrectly to all items were during analysis.

Table 3
TEST AND ITEM STATISTICS FOR DATA SET 3 (n = 797-24=773)

CLASS	RASCH	Item No.	CLASS	RASCH	CLASS	RASCH
Total Test Score	Person Ability Estimate		Item P Value	Item Diff. Estimate	Item Total Corr.	Mean Square Residual
1	-3.95	1	82.1%	-1.500	.2709	0.97
2	-3.21	2	24.8%	1.427	.2462	1.00
3	-2.74	3	14.1%	2.192	.0350	1.16
4	-2.39	4	46.3%	0.353	.3453	0.95
5	-2.11	5	49.4%	0.212	.4024	0.90
6	-1.86	6	16.8%	1.966	.1041	1.08
7	-1.64	7	28.2%	1.235	.3084	0.96
8	-1.45	8	52.3%	0.084	.2682	1.01
9	-1.26	9	64.7%	-0.492	.3291	0.95
10	-1.09	10	43.2%	0.494	.2853	0.99
11	-0.93	11	53.6%	0.026	.3177	0.97
12	-0.78	12	48.9%	0.236	.1556	1.10
13	-0.63	13	22.4%	1.578	.0945	1.10
14	-0.48	14	55.6%	-0.068	.2932	0.99
15	-0.34	15	75.3%	-1.054	.1581	1.06
16	-0.20	16	54.9%	-0.033	.2866	0.99
17	-0.07	17	22.3%	1.586	.2651	0.99
18	0.07	18	67.8%	-0.646	.2973	0.97
19	0.21	19	92.6%	-2.553	.2883	0.93
20	0.35	20	55.8%	-0.074	.4329	0.88
21	0.49	21	80.3%	-1.372	.2232	0.99
22	0.63	22	45.4%	0.394	.3451	0.95
23	0.78	23	61.2%	-0.325	.2425	1.02
24	0.94	24	42.2%	0.542	.2786	1.00
25	1.10	25	40.9%	0.602	.2996	0.98
26	1.27	26	76.1%	-1.100	.3521	0.92
27	1.45	27	85.9%	1.799	.2795	0.94
28	1.65	28	78.7%	-1.260	.2863	0.95
29	1.86	29	52.1%	0.090	.2989	0.98
30	2.10	30	40.4%	0.626	.1106	1.14
31	2.39	31	63.8%	-0.448	.2899	0.99
32	2.73	32	63.3%	-0.423	.3455	0.94
33	3.19	33	51.6%	0.113	.3171	0.97
34	3.93	34	75.4%	-1.061	.2552	0.99
		35	44.1%	0.453	.2105	1.05
r=.9931			r=.-9950		r=.-9522	

Note: Total Test Scores are based on a possible maximum of 34 points. Individuals responding correctly to all items or responding incorrectly to all items were during analysis.